Week 3 Notes

Lectures 5 & 6

University of Massachusetts Amherst, CS389

1 Multilayer Perceptrons (lecture 5)

1.1 Non-linearities

Examples of non-linearity functions can be refered below

1.1.1 Sigmoid

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1.1}$$

The derivative is:

$$\frac{d}{dx}\sigma(x) = \sigma(x) \cdot (1 - \sigma(x)) \tag{1.2}$$

1.1.2 Tanh

$$\tanh(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{1.3}$$

The derivative is:

$$\frac{d}{dx}\tanh(x) = 1 - \tanh^2(x) \tag{1.4}$$

1.1.3 ReLU

$$\text{ReLU}(x) = \max(0, x) \tag{1.5}$$

Derivative is:

$$\frac{d}{dx}\text{ReLU}(x) = \begin{cases} 0 & x < 0 \\ 1 & \text{otherwise} \end{cases} \tag{1.6}$$

2 Backpropagation (lecture 5)

As an example, we can setup a neural network with a single weight in each layer with no bias term like in Figure 1. In each layer, we pass the dot product of the weight and the input through an activation function. The output is then passed onto the next layer in the network as an input. For example, the output of the first layer is $A_1(W_1X_1)$, which then becomes the input for the second layer.

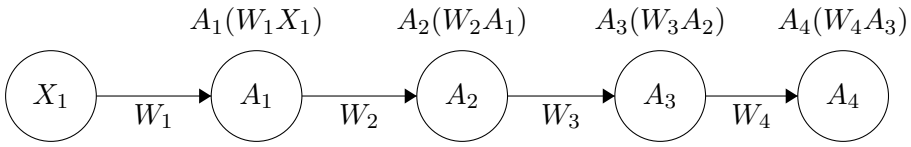

Figure 1: A 4 layer network with a single-weight each layer

where A is an activation function of any kind. Now recall the equation for the gradient descent update step as well as the gradient of loss:

$$W = W - \alpha \frac{\partial L}{\partial W} \tag{2.1}$$

$$\nabla_W \text{Loss} = \frac{\partial L}{\partial W} = \left(\frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2}, \ldots, \frac{\partial L}{\partial w_m} \right) \tag{2.2}$$

Previously, we solved this using the chain rule:

$$\frac{\partial L}{\partial W_j} = \frac{\partial \hat{y}}{\partial w_j} \frac{\partial L}{\partial \hat{y}} \tag{2.3}$$

We can extend this chain rule technique and apply it to our network in Figure 1. The idea is to consider all the terms that depend on each other in the network. The reason why we need to do this is because the input of a layer is the output of the previous layer. For example, changing w_2 will affect A_4, A_3, A_2 and L so we must consider all these variables when we try to find $\frac{\partial L}{\partial w_2}$. We can do this by multiplying all the partial derivatives along the path from loss to w_i:

$$\frac{\partial L}{\partial w_4} = \frac{\partial L}{\partial A_4} \cdot \frac{\partial A_4}{\partial w_4} \tag{2.4}$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial A_4} \cdot \frac{\partial A_4}{\partial A_3} \cdot \frac{\partial A_3}{\partial w_3} \tag{2.5}$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial A_4} \cdot \frac{\partial A_4}{\partial A_3} \cdot \frac{\partial A_3}{\partial A_2} \cdot \frac{\partial A_2}{\partial w_2} \tag{2.6}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial A_4} \cdot \frac{\partial A_4}{\partial A_3} \cdot \frac{\partial A_3}{\partial A_2} \cdot \frac{\partial A_2}{\partial A_1} \cdot \frac{\partial A_1}{\partial w_1} \tag{2.7}$$

Observe that a lot of the partial derivattives are repeated, as highlighted in blue.

2.1 Code implementation (lecture 6)

References

[1] Cooper.